DOCUMENT RESUME

ED 117 150 TM 005 004

AUTHOR Martin, Charles G.; Games, Paul A.
TITLE Selection of Subsample Sizes for the Bartlett and Kendall Test of Homogeneity of Variance.
PUB DATE [Apr 75]
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Analysis of Variance; *Hypothesis Testing; *Sampling
IDENTIFIERS Distributions (Statistical); *Homogeneity of Variance; Power (Statistics)

ABSTRACT

Power and stability of Type I error rates are investigated for the Bartlett and Kendall test of homogeneity of variance with varying subsample sizes under conditions of normality and nonnormality. The test is shown to be robust to violations of the assumption of normality when sampling is from a leptokurtic population. Suggestions for selecting subsample sizes which will produce maximum power are given for small, intermediate and large sample situations. A formula for estimating power in the equal n case is shown to give results approximating empirical results. The problem of heterogeneous within cell variances and unequal n's is discussed. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED117150

Selection of Subsample Sizes for the Bartlett and Kendall Test of Homogeneity of Variance

Charles G. Martin

U.S. Civil Service Commission

Paul A. Games

The Pennsylvania State University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 004

Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., March - April, 1975

DISCLAIMER

This paper represents research conducted at the Pennsylvania State University. It in no way reflects the views of nor obligates the U. S. Civil Service Commission.

INTRODUCTION

In research it frequently happens that K populations ($K > 2$) are to be tested for homogeneity of variance. Tests of homogeneity of variance are appropriate in two general situations. First, these tests are used when the experimenter has an a priori interest in testing the equality of variances for K independent groups. Second, they are used in testing the assumption of homogeneity of variance needed to guarantee the accuracy of certain tests on means.

Various theories in education and psychology have generated a priori hypotheses about equality of variances. For example, if high level skills build upon lower level skills as in Gagné's (1965) hierarchically arranged behaviors, small variance in the low level skills would facilitate teaching of higher skills. One of the major pieces of evidence for Cattell's two factor theory of general intelligence is the difference in test score variance between tests of fluid and tests of crystallized abilities (Cattell, 1971). In classical mental test theory (Lord and Novick, 1968) parallel forms of tests are required to have equal variances.

The use of tests of homogeneity of variance to guarantee the accuracy of tests on means is considered by many to be unnecessary because the analysis of variance is generally robust to violations of this assumption. Box (1954) and Norton (Lindquist, 1953, p. 78) have shown that this assumption is critical to the analysis of variance when n's are small or unequal. That heterogeneous variances are not important when n's are equal seems to have boundary conditions which may not have been sufficiently probed (Glass,

Peckham and Sanders, 1972). Investigations of the t statistic with unequal n's by Kohr (1970) and of three multiple comparison procedures, Multiple t test (Fisher, 1935), the Tukey Wholly Significant Difference (Tukey, 1953; Miller, 1966) and the Scheffé (Scheffé, 1953), by Howell (1971) indicate that the assumption of homogeneity of variance is critical to these procedures also.

Typically, independent random samples are compared via some function of the sample variances with a known sampling distribution when assumptions are met. Bartlett and Kendall (1946) suggested randomly dividing each of the K samples into subsamples and computing a variance estimate on each subsample. An analysis of variance (AOV) is then conducted on the variable $\ln s^2$ to test the equality of the K variances. The additive model of the AOV is met by using $\ln s^2$ as the dependent variable but not by using simply s^2. This test is described in Scheffé (1959, p. 83), Odeh and Olds (1959) and Winer (1971, p. 219). Games, Winkler and Probert (1972, p. 904) provided a description and an example including follow-up comparisons. In the present study stability of Type I error rates, power and a procedure for estimating power were investigated for the Bartlett and Kendall test with various subsample sizes when the assumption of normality was met or was violated.

TESTING EQUALITY OF VARIANCES

In addition to the Bartlett and Kendall test a wide selection of tests of homogeneity of variance is available in the statistical literature. The first approach to the variance testing problem was made by Neyman and Pearson (1931) using a likelihood-ratio statistic, approximately distributed as chi-square (K - 1 degrees of freedom). The familiar Bartlett M test is

a modification of this statistic improving the approximation to chi-square (Bartlett, 1937). A modification of the Bartlett test which adjusts M to compensate for the population kurtosis was proposed by Box and Andersen (1955). A statistic designed for the situation where just one of several populations is suspected of having a larger variance was introduced by Cochran (1941, 1951). The statistic is based on the ratio of the largest sample variance to the sum of the sample variances: $C = s^2_{max}/\Sigma s^2_k$ (Myers, 1966, p.73; Winer, 1971, p. 208). Hartley (1950) derived a statistic, F max, comparing the largest and smallest of the sample variances to reduce the computational effort typical of tests on variances (Myers, 1966, p. 73; Winer, 1971, p. 206). Cadwell (1952) suggested an extension of the Hartley technique in which variance estimates based on sums of squares are replaced by estimates based on ranges. Another test based on the analysis of variance of transformed observations was suggested by Levene (1960). Levene proposed two transformations:

$$(1)\quad Z_{ij} = |X_{ij} - \bar{X}_{.j}|$$

$$(2)\quad S_{ij} = Z^2_{ij} = (|X_{ij} - \bar{X}_{.j}|)^2.$$

Miller (1968) proposed doing an AOV on $Z'_{ij} = |X_{ij} - m_j|$ where m_j is the median of the j^{th} sample. The Foster and Burr Q test (Foster, 1964) is based on a monotone function of the coefficient of variation of the sample variances. For equal n, $Q = \Sigma s^4_k/(\Sigma s^2_k)^2$. Miller (1968) applied the Tukey jackknife technique (Mosteller and Tukey, 1968) to the variance testing problem. Layard (1973) suggested a χ^2 statistic useful with large samples. Various nonparametric tests have been developed and are discussed in Klotz (1962). Only one, the Moses test (Moses, 1963) has been used in recent comparisons with parametric tests.

An imposing problem with tests of homogeneity of variance and one of the main reasons for the proliferation of tests in this area is the general sensitivity (nonrobustness) of these statistics to violations of the normality assumption. Although the assumption is trivial to many tests on means it is crucial to tests on variances. The difference lies in the standard errors of the two statistics used when inferences are being made. For tests on means the standard error of $\overline{X} = \sigma_{\overline{X}} = \sigma_X/\sqrt{n}$ regardless of the population form but with variances the standard error of $s^2 = \sigma_{s^2} = \sigma_X^2\sqrt{\frac{2}{n-1} + \frac{\gamma_2}{n}}$ where γ_2 is the index of kurtosis for the population (Johnson and Jackson, 1959). Normal distributions have a γ_2 value of 0.0. In making inferences about s^2 the assumption of normality is necessary not only to show that the sampling distribution has a chi-square form but also to fix the magnitude of the sampling fluctuation. If the population is platykurtic ($-2 \leq \gamma_2 < 0.0$) the value of γ_2 used in the theoretical derivation will be larger than the true value and a conservative test will result. Conversely, with a leptokurtic population ($0.0 < \gamma_2 \leq +\infty$) the true value will be larger than the theoretical value, raising the probability of a Type I error, P(EI) above alpha.

Scheffé (1959, p. 337) concluded that violations of the normality assumption produce dangerous effects on inferences about variances because although the theoretical distribution may have the correct location and shape, at least for large n, it may have the wrong spread if the true γ_2 differs from zero. Box (1953) suggested that robustness is the most important characteristic of a statistic even to the extent of sacrificing power to ensure control of Type I error.

The sensitivity to nonnormality of the F test of two independent sample variances was pointed out by Pearson (1931), Geary (1947) and Gayen (1950). Box (1953) showed that this sensitivity is even greater when the number of variances exceeds two. Box showed that Bartlett's M is asymptotically distributed as $(1 + .5\gamma_2)\ \chi^2_{K-1}$. Both the F max and the Cochran tests were shown to be affected by kurtosis in much the same manner as the Bartlett test. Box carried out a small sampling study comparing the Bartlett test with the Bartlett and Kendall test for the case where the assumption of normality was violated. The results for the Bartlett test showed extremely large departures from the values expected from normal theory. In contrast, the results for the Bartlett and Kendall test gave values agreeing with what would be expected assuming that the logarithms of the variances were drawn from a normal distribution.

Levene (1960) compared the empirical sampling distributions of F ratios calculated on the Z_{ij}'s with the F distribution. Empirical probabilities were significantly different from the nominal alpha when sampling was from a double exponential distribution. Miller (1968) suggested that this condition will not necessarily improve as n increases since Z is not asymptotically distribution free. Brown and Forsythe (1974) found better agreement with the nominal alpha when deviations were taken from the median rather than the mean.

Using a monte carlo design, Miller (1968) examined the robustness of the F, Box-Andersen M', Levene S, Bartlett and Kendall (referred to as the Box test), Moses and the Tukey jackknife test for the two group case with small samples. Five distributions were used: (1) uniform, (2) normal, (3) double exponential, (4) skew double exponential, and (5) sixth power. The F test was found to be extremely nonrobust. Somewhat less sensitive

but still of questionable robustness were the Box-Andersen and jackknife tests. The Levene S, Bartlett and Kendall, and Moses tests were generally robust with empirical significance levels close to the nominally indicated levels. In terms of robustness the F, Box-Andersen and jackknife tests do not appear to be acceptable as tests of homogeneity of variance.

In a monte carlo study specifically designed to examine the robustness of tests on variances, Fellers (1972) compared the Hartley approximation to the Bartlett test (Hartley, 1940), the Cochran test, the F max test, the Bartlett and Kendall test (referred to as the Scheffé test), and three forms of the Levene test. Populations considered were normal, leptokurtic-symmetric, platykurtic-symmetric, leptokurtic-skewed and platykurtic-skewed. Equal and unequal n cases were considered for three treatment groups with total N set at 15. Fellers data showed the Bartlett test to be conservative for the platykurtic populations and extremely permissive for the leptokurtic populations. The Cochran and F max test were nonrobust for all but the near normal populations with the leptokurtic populations producing the most extreme effects. Of the three Levene tests, the S and Z lacked sufficient robustness to be considered as acceptable tests of homogeneity of variance. The third, Z', based on the absolute deviations from the median produced results so extreme as to render them uninterpretable. Only the Bartlett and Kendall test proved to be generally robust for all combinations of nonnormality and sample size.

Gartside (1972) investigated the stability of error rates when populations had normal or Weibull (leptokurtic) distributions. In addition to the Bartlett, Cochran, F max, and Bartlett and Kendall tests Gartside included modifications of the Bartlett and Bartlett and Kendall tests and added the Cadwell test. Only the Bartlett and Kendall test maintained stable error

rates for the leptokurtic population. As the number of populations being compared was increased the deviations from the nominal alpha also increased. The latter supports Box (1953) who derived this conclusion mathematically.

The question of robustness was considered in two monte carlo studies reported by Games, Winkler and Probert (1972). The first study compared the F max, Cochran, two Levene (Z and S), and the Bartlett tests. In the second study the Bartlett, Box-Andersen, Bartlett and Kendall, and Foster and Burr Q tests were compared. Samples were drawn from six populations: (1) normal, (2) slight skew, (3) moderate skew, (4) extreme skew, (5) symmetric leptokurtic and (6) rectangular. Tests were conducted on three samples of size 6 in the first study and of size 18 in the second. With n=18 two forms of the Bartlett and Kendall test were used. The first used nine subsamples of two cases each (LEV 2). The second used six subsamples of size three (LEV 3). The results indicate that the Bartlett, F max, Cochran, Levene Z and S, Foster-Burr, and Box-Andersen tests are extremely sensitive to the shape of the underlying distribution. LEV 2 was conservative for most distributions and not really sensitive to distribution form. Although LEV 3 was slightly conservative it was most robust. Games et al. concluded that on the basis of control of Type I error alone the Bartlett and Kendall test is recommended for all situations.

Only the Moses test and the Bartlett and Kendall test have been shown to be robust with respect to control of Type I error under the various conditions of nonnormality which have been investigated. These would be the recommended tests when there is suspected nonnormality in the data. Unfortunately power considerations will lead to a different recommendation.

Pearson (1966) employed a monte carlo design to compare the power of the Bartlett, F max, and Cadwell tests for five groups of five observations

each. For all situations the Bartlett test exhibited the greatest power. Pearson recommended the Bartlett test but noted that none of the tests were robust to violations of the assumption of normality.

In a monte carlo study of the two group case by Miller (1968) power for the F, Box-Andersen, Bartlett and Kendall, jackknife, Levene S and Moses tests was investigated. The most powerful test examined, the F test, was dismissed by Miller because of sensitivity to nonnormality. The Box-Andersen and the jackknife had approximately the same power and were the most powerful of the other tests. Slightly less powerful were the Bartlett and Kendall test and a form of the jackknife test. The Levene Z was not as powerful as those above. Least powerful of all the tests was the Moses test. Miller suggested using the jackknife of the Box-Andersen as a general technique for testing variances. If there is possible leptokurtosis in the population an experimenter would be sacrificing control of Type I error for power by following this suggestion.

Gartside (1972) investigated power for the Bartlett, F max, Cochran, Cadwell and Bartlett and Kendall tests for the case where the assumption of normality was met. The results showed the Bartlett test to be generally most powerful. The Cochran test was most powerful when only one of a set of variances was different. The F max and Cadwell tests showed fairly good power in all cases as did a modification of the Bartlett test. The Bartlett and Kendall test showed lower power. Gartside noted that the success of the Bartlett test must be tempered by its unstable error rates. He concluded that the more conservative Bartlett and Kendall test is preferred if there is reason to believe that the data are nonnormal.

In the two sampling studies conducted by Games, Winkler and Probert (1972) power of the selected tests on variances was investigated. Results

for the normal and extreme skew populations were similar with the Bartlett and F max tests showing nearly identical power. The Cochran and both Levene tests exhibited much lower power with the Levene S being the lowest of all. In the second study the Bartlett and Foster-Burr Q tests consistently showed the greatest power. The Bartlett and Kendall tests showed lower power for all populations with LEV 3 considerably more powerful than LEV 2. With the normal population the Box-Andersen test exhibited power near the Bartlett test but with the extreme skew population its power decreased to the level of LEV 3. Games et al. concluded that the Bartlett, Foster-Burr and F max tests are most powerful but are relatively useless for leptokurtic populations because of inflated P(EI)'s. The LEV 3 had power superior to the Box-Andersen test on the highly leptokurtic populations and stands as the best statistic for this population condition.

Games at al. (1972) noted that the biggest question in the application of the Bartlett and Kendall test is: given the number of treatments, K, and samples of n observations each, how many subsamples, m, should be used? If fewer subsamples are used the variance estimates become more stable producing a smaller mean square error for the AOV which increases power. But the degrees of freedom for the error term is also decreased when m is lowered which reduces power. There should be an optimal value for m which balances these two determinants of power.

Box (1953) first suggested an investigation to determine this optimal value of m. Miller (1968) ignored the question, claiming that the choice of subsample size rests on the shoulders of the statistician. A similar point of view was taken by Winer (1971) who stated that the number of subsamples is arbitrary. He did suggest using subsamples of approximately equal size, preferably of size larger than three. To assure reasonable power Winer

recommended having the total number of subsamples minus the number of treatment groups, i.e., the degrees of freedom of the AOV error term, at least equal to ten.

Three subsample sizes were compared for power by Gartside (1972). With n=16, subsamples of size two, four and eight were used under several conditions of variance heterogeneity with samples drawn from a normal population. Using the intermediate arrangement, i.e., four samples of size four produced the maximum power.

With n=18, Games et al. compared sampling arrangements of six subsamples of size three with nine subsamples of size two. Both arrangements provided acceptable control of Type I error although the latter was consistently conservative. Subsamples of size three produced higher power than subsamples of size two for all populations at all points where the null hypothesis was false.

Games et al. suggested using the power functions of the analysis of variance to select sample and subsample sizes for the Bartlett and Kendall test. By setting K, n and the degree of falsity of the null hypothesis approximate power could be found for different subsample sizes. Tables of the power functions of the analysis of variance are readily available (Myers, 1966; Winer, 1971). Setting K = 3 and K = 5, all n's from 12 to 36 that had any two of the numbers 3, 4, 5 and 6 as factors were explored. The noncentrality parameter of these tables, ϕ, (Myers, 1966, p. 77) is discussed later in this paper. For a highly leptokurtic population, $\gamma_2 = 6.0$, the results suggested subsamples of size three would result in maximum power up to n = 18 with little loss in power up to n = 36. Setting $\gamma_2 = 0.0$, a normal population, suggested subsamples of size three up to n = 18 but of size four from n = 18 to n = 36. Asymptotic power

theory (Miller, 1968) suggests increasing subsample size for very large n. No empirical test has been made of this method for selecting subsample size.

THE BARTLETT AND KENDALL STATISTIC

In investigations of variance heterogeneity the Bartlett and Kendall test is appropriate for a one-way or higher analysis of variance layout. For simplicity of presentation only the test for a single factor, independent groups design is discussed.

The Bartlett and Kendall test under assumptions appropriate to the analysis of variance, compares K samples of n_i ($i = 1, 2, \ldots, K$) observations each. Observations are randomized within treatment groups and divided into m_i subsamples of v_{ij} ($j = 1, 2, \ldots, m_i$) observations. On each subsample an estimate of the treatment variance, s^2_{ij}, is computed. An AOV is then conducted using $Y_{ij} = \ln s^2_{ij}$ as observations. If all v_{ij} are equal the test statistic is:

$$\frac{[\Sigma_i\, m_i(\overline{Y_{i.}} - \overline{Y_{..}})^2]\; \Sigma_j\, (m_i - 1)}{[\Sigma_i \Sigma_j\, (Y_{ij} - \overline{Y_{i.}})^2]\; (K - 1)}$$

where

$$\overline{Y_{i.}} = \Sigma_j\, Y_{ij}/m_i$$

$$\overline{Y_{..}} = \Sigma_i\, m_i \overline{Y_{i.}}/\Sigma_i\, m_i \quad .$$

A weighted least squares solution was provided for the situation where subsample sizes (v_{ij}) are not all equal by Scheffé (1959, p. 85). The test statistic for unequal v_{ij} is:

$$\frac{[\Sigma_i \overline{u_{i.}}(\overline{X_{i.}} - \overline{X_{..}})^2]\ \Sigma_i\ (m_i - 1)}{[\Sigma_i \Sigma_j\ u_{ij}(Y_{ij} - \overline{X_{i.}})^2]\ (K - 1)}$$

where

$$u_{ij} = v_{ij} - 1$$
$$\overline{u_{i.}} = \Sigma_j\ u_{ij}$$
$$\overline{u_{..}} = \Sigma_i\ \overline{u_{i.}} = \Sigma_i \Sigma_j\ u_{ij}$$
$$\overline{X_{i.}} = \Sigma_j\ u_{ij} Y_{ij} / \overline{u_{i.}}$$
$$\overline{X_{..}} = \Sigma_i \Sigma_j\ u_{ij} Y_{ij} / \overline{u_{..}}\ .$$

These two statistics can be compared to the F distribution with K - 1 and $\Sigma_i\ (m_i - 1)$ degrees of freedom.

The power functions of the analysis of variance appear apropriate for estimating power and for selecting n and v (all v_{ij} equal) for a given power for the Bartlett and Kendall test. A priori estimates of power can be made using the Pearson and Hartley (1951, p. 112) charts of the power functions for the analysis of variance. Scheffé (1959, p. 85) showed that $\sigma^2_{\ln s^2} \approx \frac{2}{v - 1} + \frac{\gamma_2}{v}$ which is the $E(MS_W)$ in an AOV on $\ln s^2$. Let ϕ be the noncentrality parameter of these charts (Myers, 1966, p. 77). Given K sets of m independent observations each, each observation being the logarithmic transformation of a sample variance of v observations

$$\phi = \sqrt{\frac{m\theta}{\sigma^2_{\ln s^2}}} \approx \sqrt{\frac{m\theta}{\frac{2}{v - 1} + \frac{\gamma_2}{v}}}$$

where

$$\theta = \Sigma_i\ (\ln \sigma_i^2 - \overline{\ln \sigma^2})^2 / K$$
$$\overline{\ln \sigma^2} = \Sigma_i\ \ln \sigma_i^2 / K\ .$$

METHOD

Because of the intractability of using direct mathematical analysis a monte carlo design was employed. For each combination of conditions investigated a simulated analysis was repeated 1,000 times in four blocks of 250 times each. The number and proportion of rejections of the null hypothesis at the one and five percent levels were recorded for each of the four blocks. The frequencies of rejection became dependent measures in analyses of variance of the conditions investigated.

In this study only the three independent group, equal n case for the Bartlett and Kendall test was investigated. Sample sizes, representing small, intermediate and large sample situations with n's of 18, 36 and 48 respectively were selected. For each n a set of six v's was investigated: n = 18, v = 2, 3, 4, 5, 6, 7; n = 36, v = 3, 4, 5, 6, 7, 8; n = 48, v = 5, 6, 7, 8, 9, 10. It was predicted that from each set one v would be found which would produce maximum power for the given sample size. Where n/v was not an integer v is the minimum subsample size. In this case the last subsample formed consisted of v plus any remaining elements (always less than v).

Control of Type I error was investigated for the case of equal treatment variances. To investigate power three conditions of variance heterogeneity were used. Variances were formed by multiplying each element in a treatment group by a constant representing the desired standard deviation for that group. Variances were chosen for each sample size which would provide diverse power points over the complete power range. Table 1 presents the constants selected and θ, a measure of the degree of variability in the set of variances.

The simulations consisted of a series of experiments, each conducted with three samples, representing three treatment groups, of n observations each. Sampling was random with replacement from two populations of 10,000 cases. A normally distributed population and a population with extreme skewness and leptokurtosis (χ^2 with 2 df) were used for the situations where the normality assumption was met and violated respectively. The normally distributed population was constructed by dividing it into 28 intervals of 0.3 standard deviation units ranging from -4.2 to 4.2 standard deviations with known expected proportions of cases for each interval. Uniform values of half the interval width were randomly generated and added to or subtracted from the interval midpoint to form normal deviates. The leptokurtic distribution was constructed by forming cases of the sum of two squared normal deviates. A procedure by Chen (1971) (made available after the study using a normal population had been conducted) was used to generate these deviates. Parameters (μ, σ^2, γ_1, γ_2) were computed for each population. The parameters were $\mu = 0.0003$, $\sigma^2 = 1.0123$, $\gamma_1 = -0.0010$, $\gamma_2 = -0.0166$ and $\mu = 1.9701$, $\sigma^2 = 4.0290$, $\gamma_1 = 2.0275$, $\gamma_2 = 6.0105$ for the normal and leptokurtic populations respectively.

A priori estimates of power using ϕ (Myers, 1966, p. 77) were compared with empirical results to examine the accuracy of this approximation. Phi was computed as if v's in a given sample were actually equal. Values of ϕ were calculated and approximate powers were taken from power curves in Myers (1966, p. 390). Empirical powers were the proportions of rejection of a false null hypothesis.

RESULTS

Analysis of Type I errors was conducted in terms of proportions of rejection (p) of a true null hypothesis. Of 72 sample p's only three differed significantly at the .05 level from the nominal error rate. Only one of these significant p's occurred with the leptokurtic population. Table 2 presents sample p's for both populations at the one and five percent levels. These p's are proportions averaged over the four blocks of trials. The Bartlett and Kendall test was shown to be robust when populations are extremely leptokurtic for all combinations of n and v investigated.

The focus of the present study was on selecting subsample sizes which would produce maximum power for a given n. To compare power produced by the different v's, analyses of variance were conducted for the conditions investigated. The dependent measures were the frequencies of rejection for the four blocks of trials. Subsample size, population form, and θ were the factors considered. The three main effects and the population form by subsample size interaction were found to be significant ($\alpha = .05$). The θ effect was expected and is trivial. It is simply a measure of the deviation from the null hypothesis. A population form effect was expected and indicates a lower power for the leptokurtic population.

The interaction of population form and subsample size represents the major complication of the study. It indicates that a single v may not produce maximum power for both population forms. Had the desired results occurred there would have been only a main subsample size effect and no interaction. Because of the interaction the subsample size effect was

investigated for each population form separately. Table 3 presents the mean frequencies of rejections (averaged over the three θ conditions and four blocks of trials) for each v under each population form at the one and five percent levels. The v's are ranked in terms of power produced. Multiple comparisons via the Newman-Keuls technique indicated that in most cases no single v was optimal (produced the greatest power) in any given set. When the mean power was significantly less ($\alpha = .05$) than the maximum in the column it is marked with an *.

Considering both nominal error rates for $n = 18$, optimal v's were $v = 4$ for the normal population and $v = 3$ for the leptokurtic population. At $n = 36$ with the normal population, no significant differences were found between v's of six and seven. The leptokurtic results provided no clear way to choose between $v = 4$, $v = 5$, or $v = 6$. Thus with $n = 36$, one could generally use $v = 6$ with relatively little danger of appreciable power loss. When $n = 48$ the results were clearer. For the normal population $v = 8$ was most powerful and for the leptokurtic population $v = 6$ and $v = 5$ produced approximately equal maximum power.

A priori estimates of power based on ϕ (Myers, 1966, p. 77) were compared with power empirically produced in the sampling study. A moderate degree of correspondence was found between the two. Pearson r's were computed for each combination of n, population form and nominal error rate. The minimum r computed was .86 and the maximum was .98 over the twelve sets of data. A large part of this relationship was due to the differences in θ. As an example Table 4 presents empirical and a priori powers for $n = 48$ and $\alpha = .05$. The AOV power functions appear useful in providing rough estimates of power for the Bartlett and Kendall test.

DISCUSSION

The robustness of the Bartlett and Kendall test was supported. Even when the population from which samples are drawn is extremely leptokurtic the Bartlett and Kendall test provides control of P(EI) at the nominal alpha level. Control is unaffected by sample and subsample sizes. This test may be recommended for general use even when populations have suspected leptokurtosis.

Several researchers (Box, 1953; Games et al., 1972; Gartside, 1972; Miller, 1968; Scheffé, 1959) have pointed out the lack of knowledge for choosing v in an analysis. This study suggests two rules of thumb for selecting a most powerful v for a fixed n. Find a value near $\sqrt{n}$, rounding noninteger values higher if the subject population is normal or lower if leptokurtosis is suspected. A second suggestion is to choose values at this point which are even divisors of n. The power differences between adjacent values of v are sufficiently small so that there usually would be little power loss if the value used differs by only one from the optimum value of v. These rules are in agreement with Gartside (1972) who noted that an intermediate arrangement would appear to give more power.

The given approximation to ϕ and the power functions of the analysis of variance are shown to be generally accurate for selecting power for the Bartlett and Kendall test. Using the above rules of thumb for selection of v, an experimenter should be able to approximate the power for any n chosen.

A problem arises if v's are selected in the suggested manner when n's are unequal. In this situation the v's will also be unequal. The treatment group with the largest n and thus largest v will have the most stable

variance estimates. Hence this treatment group will have a smaller within cell variance in the AOV. Conversely, the group with the smallest n will have a larger within cell variance. Since m = n/v the sample sizes for the AOV will also be unequal. This combination of heterogeneous within cell variances and unequal n's may present problems for the AOV.

REFERENCES

Bartlett, M. S. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London, 1937, 160, 268-282.

Bartlett, M. S. and Kendall, D. G. The statistical analysis of variance-heterogeneity and the logarithmic transformation. Journal of the Royal Statistical Society, Supplement 8, 1946, 128-138.

Box, G. E. P. Non-normality and tests of variances. Biometrika, 1953, 40, 318-335.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one way classification. Annals of Mathematical Statistics, 1954, 25, 290-302.

Box, G. E. P. and Andersen, R. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, 1955, 17, 1-26.

Brown, M. B. and Forsythe, A. B. Robust tests for the equality of variances. Journal of the American Statistical Association, 1974, 69, 364-367.

Cadwell, J. H. Approximating the distributions of measures of dispersion by a power of χ^2. Biometrika, 1953, 40, 336-346.

Cattell, R. B. Abilities: Their structure, growth, and action. New York: Houghton-Mifflin,,1971.

Chen, E. H. Random normal number generator for 32-bit-word computers. Journal of the American Statistical Association, 1971, 66, 400-403.

Cochran, W. G. The distribution of the largest of a set of estimated

variances as a fraction of their total. Annals of Eugenics, 1941, 11, 47-52.

Cochran, W. G. Testing a linear relation among variances. Biometrics, 1951, 7, 17-32.

Fellers, R. R. The effects of nonnormality and sample size on the robustness of tests of homogeneity of variance. Paper presented at the Northeast Educational Research Association Meeting, 1972.

Fisher, R. A. The design of experiments. Edinburgh: Oliver & Boyd, 1935.

Foster, L. A. The Q-test for equality of variances. Unpublished doctoral dissertation, Purdue University. Ann Arbor, Michigan: University Microfilms, 1964, No. 65-5008.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart & Winston, Inc., 1965.

Games, P. A., Winkler, H. B. and Probert, D. A. Robust tests for homogeneity of variance. Educational and Psychological Measurement, 1972, 32, 887-909.

Gartside, P.S. A study of methods for comparing several variances. Journal of the American Statistical Association, 1972, 67, 342-346.

Gayen, A. K. The distribution of the variance ratio in random samples of any given size drawn from non-normal populations. Biometrika, 1950, 37, 236-255.

Geary, R. C. Testing for nonnormality, Biometrika, 1947, 34, 209.

Glass, G. V., Peckham, P. D. and Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 1972, 42, 237-288.

Hartley, H. O. Testing the homogeneity of a set of variances. Biometrika, 1940, 31, 249-255.

Hartley, H. O. The maximum F-ratio as a short-cut test for heterogeneity of variance. Biometrika, 1950, 37, 308-312.

Howell, J. F. The effects of variance heterogeneity on selected multiple comparison procedures. Unpublished doctoral dissertation, The Pennsylvania State University, 1971.

Johnson, P. O. and Jackson, R. W. B. Modern statistical methods: Descriptive and inductive. Chicago, Ill.: Rand McNally, 1959.

Klotz, J. Nonparametric tests for scale. Annals of Mathematical Statistics, 1962, 33, 498-512.

Kohr, R. L. A comparison of statistical procedures for testing $\mu_1 = \mu_2$ with unequal n's and variances. Unpublished doctoral dissertation, The Pennsylvania State University, 1970.

Layard, M. W. J. Robust large sample tests for homogeneity of variances. Journal of the American Statistical Association, 1973, 68, 195-198.

Levene, H. Robust test for equality of variances. In I. Olkin (ed.) Contributions to probability and statistics. Stanford, Calif.: Stanford University Press, 1960, 278-292.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston, Houghton-Mifflin, 1953.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Miller, R. G., Jr. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Miller, R. G., Jr. Jackknifing variances. Annals of Mathematical Statistics, 1968, 39, 567-582.

Moses, L. E. Rank tests of dispersion. Annals of Mathematical Statistics, 1963, 35, 1594-1605.

Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronson, The handbook of social psychology, (2nd ed.) Reading, Mass.: Addison-Wesley, 1968, 80-203.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Neyman, J. and Pearson, E. S. On the problem of k samples. Bulletin de l'Academie Polonaise des Sciences et des Lettres, A 1931, 460-481.

Odeh, R. I. and Olds, E. G. Notes on the two analysis of variance of logarithms of variances. WADC tech. Note 59-82, ASTIA doc. No. AD211917, 1959, OTS, U. S. Dept. of Commerce, Wshington, D. C.

Pearson, E. S. The analysis of variance in cases of non-normal variation. Biometrika, 1931, 23, 114-133.

Pearson, E. S. Alternative tests for heterogeneity of variance: Some monte carlo results. Biometrika, 1966, 53, 229-234.

Pearson, E. S. and Hartley, H. O. Charts of the power function for analysis of variance tests derived from the noncentral F-distribution. Biometrika, 1951, 38, 112-130.

Scheffé, H. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Tukey, J. W. The problem of multiple comparisons. Unpublished manuscipt, Princeton University, 1953.

Winer, B. G. Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill, 1971.

TABLE 1

Constants and θ's

Sample Size	Variance Condition	θ*	Group 1	Group 2	Group 3
18	1	0.0000	1	1	1
18	2	0.0806	1	2	2
18	3	0.1552	1	2	3
18	4	0.2417	1	2	4
36	1	0.0000	1	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$
36	2	0.0388	1	$\frac{1}{\sqrt{2}}$	2
36	3	0.0604	1	2	2
36	4	0.0806	1	2	2
48	1	0.0000	1	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$
48	2	0.0201	1	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$
48	3	0.0388	1	$\frac{1}{\sqrt{2}}$	2
48	4	0.0604	1	$\frac{1}{\sqrt{2}}$	2

* Theta's reported in the present study are based on common logarithms.

TABLE 2

Proportion of Significant Results

When H_0 is True

n	v	Skewed-Leptokurtic Population α=.01	Skewed-Leptokurtic Population α=.05	Normal Population α=.01	Normal Population α=.05
18	2	.010	.051	.006	.034*
18	3	.014	.058	.010	.041
18	4	.008	.051	.015	.052
18	5	.008	.065*	.006	.044
18	6	.006	.049	.007	.051
18	7	.009	.050	.012	.057
36	3	.013	.049	.004	.041
36	4	.013	.052	.009	.045
36	5	.006	.061	.008	.042
36	6	.007	.039	.005	.030*
36	7	.011	.053	.009	.047
36	8	.008	.052	.010	.045
48	5	.013	.044	.013	.051
48	6	.011	.044	.014	.054
48	7	.005	.040	.013	.049
48	8	.006	.048	.010	.056
48	9	.012	.053	.00[a]	.047
48	10	.009	.040	.010	.059

* Represents significant deviation from α.

TABLE 3

Mean Frequency of Rejection for
Population Form X Subsample
Size Interaction

Normal Population				Skewed-Leptokurtic Population			
			n = 18				
v	α=.01	v	α=.05	v	α=.01	v	α=.05
4	111.25	6	181.67	3	52.17	3	110.83
3	104.33*	4	177.67	4	46.92*	4	108.25
6	99.58*	5	176.83	2	41.50*	6	102.92*
5	92.25*	3	167.83*	6	37.50*	5	98.08*
2	62.75*	7	133.17*	5	32.58*	2	91.00*
7	46.42*	2	115.58*	7	17.25*	7	68.58*
			n = 36				
v	α=.01	v	α=.05	v	α=.01	v	α=.05
7	127.75	7	197.25	4	49.33	5	105.58
5	125.17	6	194.00	3	47.67	6	104.17
6	122.50*	8	190.25*	5	45.83	4	102.58
4	118.58*	5	187.92*	6	44.58	7	97.50*
8	113.42*	4	175.83*	7	41.33*	3	94.58*
3	94.92*	3	152.83*	8	36.83*	8	92.75*
			n = 48				
v	α=.01	v	α=.05	v	α=.01	v	α=.05
8	126.42	8	184.83	6	42.83	6	92.75
6	121.42*	9	182.75	5	41.75	5	90.17
9	120.92*	6	175.58*	7	37.00*	7	88.17
5	117.83*	7	174.58*	8	36.08*	8	84.58*
7	116.25*	10	173.08*	9	36.00*	9	82.83*
10	104.67*	5	171.58*	10	28.42*	10	73.42*

* Represents significant deviation from maximum, α=.05.

TABLE 4

Empirical and A Priori Powers
$n = 48, \ \alpha = .05$

v	df_E	θ	Normal Population A Priori ϕ	Normal Population Estimated	Normal Population Obtained	Skewed-Leptokurtic Population A Priori ϕ	Skewed-Leptokurtic Population Estimated	Skewed-Leptokurtic Population Obtained
5	24	.0201	1.386	.52	.443	0.752	.21	.204
5	24	.0388	1.924	.80	.720	1.044	.30	.354
5	24	.0604	2.401	.94	.896	1.302	.42	.524
6	21	.0201	1.461	.51	.462	0.781	.21	.217
6	21	.0388	2.029	.82	.729	1.084	.40	.359
6	21	.0604	2.531	.95	.916	1.353	.48	.537
7	15	.0201	1.386	.50	.448	0.734	.20	.214
7	15	.0388	1.924	.76	.734	1.018	.29	.340
7	15	.0604	2.401	.91	.913	1.271	.40	.504
8	15	.0201	1.497	.55	.499	0.786	.19	.184
8	15	.0388	2.079	.80	.781	1.092	.31	.338
8	15	.0604	2.594	.96	.938	1.362	.48	.493
9	12	.0201	1.461	.51	.491	0.763	.15	.188
9	12	.0388	2.029	.78	.777	1.059	.27	.339
9	12	.0604	2.531	.94	.925	1.322	.40	.467
10	9	.0201	1.386	.41	.454	0.721	.15	.177
10	9	.0388	1.924	.70	.720	1.000	.24	.286
10	9	.0604	2.401	.89	.903	1.248	.33	.418

APPENDIX

Computational Example of the Bartlett and Kendall Test
n_k=7,7,7, v_k=3,3,3

Data Matrix

X_{i1}	s^2_{i1}	ln s^2_{i1}	X_{i2}	s^2_{i2}	ln s^2_{i2}	X_{i3}	s^2_{i3}	ln s^2_{i3}
14			10			31		
8	9.0000	2.1972	9	19.0000	2.9444	10	120.3333	4.7903
11			17			15		
10			12			15		
14	2.9167	1.0704	12	10.2499	2.3272	36	94.2500	4.5459
12			18			24		
11			17			16		
	s^2_1 = 4.6190			s^2_2 = 13.6190			s^2_3 = 92.0000	

Analysis of Variance

Y_{ik} = ln s^2_{ik} entries for ANOVA

T_1	T_2	T_3	
2.1972	2.9444	4.7903	
1.0704	2.3272	4.5459	
1.6338	2.6358	4.6681	$\overline{Y_k}$
5.1233	13.6358	106.4952	Antilog $(\overline{Y_k})$

$\Sigma\Sigma\ Y$ = 17.8756
$\Sigma\Sigma\ Y^2$ = 66.6719
SS_{TOT} = 10.4157
SS_W = 0.8551
SS_{BET} = 9.5606

$$F = \frac{9.5606 / 2}{.8551 / 3} = \frac{4.7803}{.2850} =$$

16.7708 with df = 2,3